Genome-wide association analysis of plasma lipidome identifies 495 genetic associations

Linda Ottensmann[1], Rubina Tabassum[1], Sanni E. Ruotsalainen[1], Mathias J. Gerl[2], Christian Klose[2], Elisabeth Widén[1], FinnGen[6], Kai Simons[2], Samuli Ripatti[1,3,4], Matti Pirinen[1,3,5]
[1]Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Helsinki, Finland; [2]Lipotype GmbH, Dresden, Germany; [3]Department of Public Health, Clinicum, Faculty of Medicine, University of Helsinki, Helsinki, Finland; [4]Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA; [5]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Abstract

Human plasma lipidome captures risk for cardio-metabolic diseases. To discover new lipid-associated variants and understand link between lipid species and cardiometabolic disorders, we performed univariate and multivariate genome-wide analyses of 179 lipid species in 7,174 Finnish individuals. We further fine-mapped the associated loci, prioritized genes, and examined their disease links in 377,277 FinnGen participants. We identified 495 genome-trait associations in 56 genetic loci including 9 novel loci, with a considerable boost provided by multivariate analysis. For 26 loci, fine-mapping identified variants with a high causal probability, including 14 coding variants indicating likely causal genes. Phenome-wide analysis across 953 disease endpoints in FinnGen revealed disease associations for 40 lipid loci. For 11 known coronary artery disease risk variants, we detected strong associations with lipid species. Our study demonstrates the power of multivariate genetic analysis in correlated lipidomics data and reveals genetic links between diseases and detailed lipid measures beyond standard lipids.

Introduction

Cardiovascular disease (CVD) is the leading cause of mortality and morbidity worldwide [1] with an estimated heritability of about 50% [2]. Plasma lipids, routinely measured via high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides (TG), and total cholesterol (TC), are established risk factors for CVD [3]. The modern efficient lipidomics technologies have extended considerably our understanding of the variability and width of circulating lipids. Lipid species including, for example, Cholesterol esters (CE), Ceramides (CER), Diacylglycerols (DAG), Lysophosphatidylcholines (LPC), Phosphatidylcholines (PC), Phosphatidylcholine-ether (PCO), Phosphatidylethanolamines (PE), Phosphatidylethanolamine-ethers (PEO), Sphingomyelins (SM) and Triacylglycerols (TAG) potentially improve CVD risk assessment over standard lipids [4]–[14]. Eventually, a better understanding of biological factors underlying lipid metabolism and its connection to CVD pathophysiology may also provide new treatment options for CVD.

Genome-wide association studies (GWAS) have revolutionized our understanding of genetic variation behind lipid levels [15]–[35]. With growing sample sizes, more efficient genetic fine-mapping methods, and the use of population isolates like Finland, several likely causal coding variants and genes have been identified. For example, recently reported stop-gained variants in *CD36*, *ANGPTL8*, and *PDE3B* provide potential targets for the next generation of lipid-lowering medications [20], [36].

Very large genetic studies have already been conducted for the standard lipids. For example, the multi-ethnic meta-analysis from the Million Veteran Program study, with a sample size of > 600,000 participants, identified 306 loci associated with the standard lipids [19] and a multi-ethnic meta-analysis in 1.65 million individuals identified 941 loci [20]. Despite much smaller sample sizes of lipidome GWAS, they have identified new lipid-associated genetic variants and provided insights into the genetic architecture of lipid metabolism and cardiometabolic diseases. Additionally, the high-dimensional and correlated structure of the lipidome [27] can be utilized in a multivariate framework [28] to increase statistical power to identify new genetic associations but, to our knowledge, such analyses have not been reported so far.

Here we report univariate and multivariate GWAS of 179 lipid species from 13 lipid classes in 7,174 Finnish individuals from the GeneRISK cohort, followed by a phenome-wide association study (PheWAS) of the identified lipid-associated genetic loci in 377,277 biobank participants of the FinnGen study. Altogether, we identified 56 lipid-associated loci including 9 new loci, 2 of which were identified in univariate GWAS (*AGPAT2*, *SGPL1)*, and 7 were only revealed through multivariate GWAS (*DTL*, *STK39*, *CDS1*, *KCNJ12*, *YPEL2*, *SPHK2*, *AGPAT3*) demonstrating the gain in statistical power provided by multivariate techniques. Fine-mapping identified variants with high causal probabilities for 26 loci. We also present detailed lipidomic profiles of known CAD-associated variants. Through the large genome-wide investigation of lipidomic measurements and a new multivariate approach, we provide new lipid-associated loci and new insights into lipid metabolism.

Results

Using shotgun lipidomics, we detected 179 lipid species belonging to 13 lipid classes covering 4 major lipid categories: glycerolipids, glycerophospholipids, sphingolipids, and sterols (Figure 1, Supplementary Table 2). Hierarchical clustering based on absolute pairwise Pearson correlations of plasma levels of lipids revealed 11 clusters of correlated lipids (Figure 1, Supplementary Figures 10-11), which were used for multivariate GWAS. Lipid species included in each cluster are listed in Supplementary Table 1, and the pairwise correlations between lipid species in each cluster are provided in Supplementary Figure 12 and Supplementary Table 18.

Heritability of lipid species

We estimated the SNP-based heritability of all 179 lipid species using >849k high-quality independent genetic variants. The heritability estimates ranged from 0.0 to 0.45 (Figure 2 and Supplementary Table 2). Sphingomyelins (SMs) had the highest estimated median heritability (median=0.35, range=0.18-0.40) followed by Ceramides (Cer) (median=0.34, range=0.05-0.36). Phosphatidylcholine-ethers (PCO) showed the smallest median heritability (median = 0.12, range = 0–0.32) preceded by Phosphatidylethanolamines (PEO) (median = 0.13, range = 0.08–0.14). Lipids containing long-chain polyunsaturated fatty acids (PUFA) (C20:4, C20:5, and C22:6 acyl chains) had slightly higher median heritability (median=0.27, range=0-0.45) compared with other lipid species (median=0.23, range=0-0.40). PC 18:0;0_20:4;0 had the highest heritability (0.45, SE=0.05) of all lipid species followed by CE 20:4;0 (0.44, SE=0.05). The heritability estimates for lipid species grouped by lipid classes, lipid categories, and PUFA acyl chains are shown in box plots in Supplementary Figure 1.

Univariate and multivariate GWAS

We performed univariate GWAS for 179 lipid species and multivariate GWAS for 11 clusters using ~11.3M high-quality genetic variants with minor-allele frequency (MAF) > 0.002. In the univariate GWAS of 179 lipid species, we identified 26,969 variant-lipid associations at the Bonferroni-corrected significance (BFS) threshold (P<7.35e-10) after correction for 68 principal components explaining 90% of the variance in lipidome. The multivariate GWAS of 11 clusters revealed 13,157 variant-cluster associations at BFS (P<4.55e-9). Genomic inflation factors for univariate and multivariate GWAS ranged between 0.99 and 1.14 (Supplementary Table 5). Manhattan plots for lipid classes and multivariate analyses are shown in Supplementary Data 1.

To define independent loci across the lipid species and clusters, we first identified lead variants, individually for each univariate (N=179) and multivariate (N=11) trait, iteratively as the variant with the lowest *P*-value. Then the ±1.5Mb regions around the lead variants were defined as lipid-associated genomic regions (GWAS regions). A total of 495 BFS GWAS regions (357 and 138 from univariate and multivariate GWAS respectively) were identified. We identified a set of most probable causal variants in each GWAS region through fine-mapping and considered each of them as representing a single association signal. We merged the identified signals that were in linkage disequilibrium (LD; $r^2 \geq 0.1$) and combined the overlapping regions across all 190 GWAS to form a non-overlapping set of lipid-associated loci. Through this process, described in detail in Methods, Supplementary Figure 3, and Supplementary Figure 4 for the locus *LPL*, we identified 98 signals (Supplementary Data 2) located in 56 independent loci across all 190 GWAS traits (Supplementary Figure 2, Table 1). The number of associated loci per lipid species correlated positively with estimated heritability (adjusted r^2=0.3125, *P*=2.5e-16), (Supplementary Figure 5). We identified 29 additional loci that were associated with lipid species or lipid clusters at genome-wide significance (GWS) but did not reach BFS (Supplementary Table 3).

Table 1. Loci reaching the Bonferroni-corrected significance level for univariate (uv) or multivariate (mv) GWAS. Cluster number is given as mv trait and lipid species name as uv trait. Loci found by previous studies for standard lipids or lipid species (LS) are marked with an X. minP = minimum of *P*-values of multivariate GWAS or univariate GWAS. Novel loci are bolded. Variants are named by chromosome:base pair position (GRCh38).

Locus	Mv trait	Mv lead variant	Mv minP	Uv trait	Uv lead variant	Uv minP	HDL	LDL	TC	TG	LS
RF00019	4	1:39937698	1e-12	LPE 18:0;0	1:39937698	4e-13				X	X
PCSK9	7	1:55039974	3e-23	SM 34:1;2	1:55039974	2e-19		X	X		X
DOCK7	2	1:62455915	4e-28	PI 18:0;0_20:4;0	1:62662654	6e-25	X	X	X	X	X
AC105942.1	5	1:94928979	1e-25	PC 18:0;0_18:2;0	1:95232272	7e-13					X
DTL	5	1:212081294	4e-11	PC 18:0;0_18:2;0	1:212081294	3e-5					
MARC1	1	1:220800221	4e-8	TAG 54:4;0	1:220800221	5e-11	X	X	X	X	
GALNT2	10	1:230167766	2e-9	PC O-16:1;0/18:1;0	1:230167766	4e-10	X			X	X
APOB	7	2:21041028	4e-11	CE 16:0;0	2:21041028	3e-9	X	X	X	X	X
GCKR	3	2:27508073	1e-17	TAG 50:4;0	2:27508073	4e-22	X	X	X	X	X
ABCG8	9	2:43847292	3e-32	CE 20:2;0	2:43847292	9e-32		X	X	X	X
STK39	8	2:168292271	3e-10	SM 34:2;2	2:168292271	1e-5					
AC021074.3	2	3:142936448	1e-39	PI 18:0;0_18:1;0	3:142936448	5e-20		X	X	X	X
ANKRD17	8	4:73167847	4e-15	SM 40:1;2	4:73167847	3e-17	X	X	X	X	
CDS1	2	4:84647685	2e-9	PI 16:0;0_18:2;0	4:84647685	3e-5					
ELOVL6	3	4:110207431	5e-33	CE 16:1;0	4:110207431	3e-7	X			X	
SMIM13	6	6:11089522	6e-20	CE 22:6;0	6:11089522	2e-5					X
AGPAT1	3	6:32168770	4e-31	TAG 50:1;0	6:32168770	4e-9	X	X	X	X	X
PEX6	6	6:42979275	5e-13	PC 18:0;0_22:6;0	6:42963486	3e-9	X	X		X	
NPC1L1	8	7:44541277	9e-11	CE 18:0;0	7:44542387	4e-8		X	X		
MLXIPL	8	7:73606007	9e-12	DAG 18:1;0_18:2;0	7:73599571	9e-11	X		X	X	X
AC022784.1	3	8:9326154	4e-13	PC 18:0;0_18:2;0	8:9326154	9e-8	X	X	X	X	
LPL	1	8:19967357	2e-10	TAG 56:6;0	8:19970337	2e-12	X			X	X
ERMP1	8	9:5612441	1e-20	SM 32:1;2	9:5811257	1e-8		X	X		
ABO	8	9:133266456	2e-10	CE 18:0;0	9:133273983	7e-14	X	X	X		X
AGPAT2	3	9:136677616	2e-12	PC 16:0;0_22:5;0	9:136677616	4e-12					
JMJD1C	8	10:63364338	4e-9	Cer 42:2;2	10:63364338	4e-6	X			X	X
SGPL1	8	10:70843134	6e-17	Cer 42:2;2	10:70843134	2e-10					
PKD2L1	3	10:100315722	4e-26	PC 16:1;0_18:1;0	10:100315722	1e-11		X			X
GPAM	2	10:112190660	1e-13	PI 18:0;0_20:4;0	10:112190660	1e-5	X	X	X	X	
PNLIPRP2	3	10:116638373	9e-12	PC 16:0;0_18:2;0	10:116638373	1e-6	X				X
MYRF	7	11:61776027	<5e-324	PC 18:0;0_20:4;0	11:61785208	<5e-324	X	X	X	X	X
CPT1A	3	11:68794860	9e-18	CE 20:4;0	11:68794860	7e-11	X			X	
RN7SL786P	3	11:75745010	2e-14	PC 18:1;0_20:2;0	11:75734293	4e-9	X		X		X
ZPR1	3	11:116778201	3e-42	TAG 54:4;0	11:116778201	9e-39	X	X	X	X	X
SOAT2	8	12:53112581	7e-35	CE 18:0;0	12:53118972	4e-24					X
HNF1A	8	12:121000508	6e-22	SM 38:2;2	12:121000508	4e-12	X	X	X		X
AL161670.1	8	14:63768838	2e-197	SM 32:1;2	14:63768838	3e-95		X	X	X	X
LIPC	2	15:58386313	2e-126	PE 16:0;0_20:4;0	15:58388755	4e-104	X	X	X	X	X
NTAN1	3	16:15038117	1e-52	CE 20:3;0	16:15038105	2e-36	X			X	X
CETP	3	16:56960616	1e-33	PC 16:0;0_18:2;0	16:56960616	8e-15	X	X	X	X	X
LCAT	7	16:67942417	1e-16	CE 20:4;0	16:67942417	2e-4	X		X		
GLTPD2	8	17:4789345	1e-79	SM 40:1;2	17:4789345	4e-60			X	X	X
KCNJ12	7	17:21386711	2e-9	PC 16:1;0_20:4;0	17:21386711	2e-6					
YPEL2	11	17:59341010	2e-9	PC O-16:0;0/20:4;0	17:59341010	7e-8					
ABHD3	2	18:21651694	5e-38	PC 14:0;0_18:2;0	18:21651694	7e-20					X
SMUG1P1	2	18:49656294	1e-15	PI 18:1;0_18:1;0	18:49656294	8e-17	X	X	X		X
CERS4	8	19:8209156	2e-189	SM 38:2;2	19:8209156	6e-53	X	X	X	X	X
TM6SF2	8	19:19141970	3e-24	TAG 56:6;0	19:19485324	2e-15		X	X	X	X
APOE	8	19:44908822	2e-65	CE 18:2;0	19:44908822	8e-36	X	X	X	X	X
SPHK2	8	19:48629610	6e-16	SM 34:2;2	19:48629610	1e-7					
TMC4	2	19:54173495	2e-301	PI 18:0;0_18:2;0	19:54173495	5e-107		X			X
LINC01722	8	20:12978039	1e-135	Cer 42:2;2	20:12982070	7e-52		X	X		X
NINL	3	20:25482746	6e-12	CE 20:3;0	20:25482746	2e-6					X
HNF4A	2	20:44413724	1e-10	CE 18:3;0	20:44413724	1e-10	X	X	X		
AGPAT3	3	21:43971391	7e-17	PC 16:0;0_22:5;0	21:43971391	5e-9					
PNPLA3	1	22:43945024	3e-9	TAG 56:6;0	22:43928847	6e-19	X	X	X	X	X

Further, for each of the multivariate associations, MetaPhat [37] was applied to identify the traits driving the multivariate association. For 61 of all 138 multivariate BFS GWAS regions, a single trait was identified to be driving the association and for 47 of the regions, 2 to 3 driver traits were identified. The driver traits and other MetaPhat results for associations reaching GWS in the multivariate analysis are listed in Supplementary Table 4.

Next, we compared the findings of the univariate GWAS and multivariate GWAS. Of the 138 BFS GWAS regions identified in multivariate analysis, 55 regions did not reach BFS in any univariate analysis of the traits included in that multivariate analysis, for any variant in LD with the lead variant ($r^2 > 0.1$) of the multivariate analysis. The multivariate analysis identified 21 loci not found by the univariate analysis. A comparison of the *P*-values of the lead variants in the 56 loci in the univariate and multivariate GWAS showed that all the loci identified by univariate GWAS reached BFS in the multivariate GWAS, except *MARC1* which only reached GWS (Figure 3). We observed much lower *P*-values in univariate compared to multivariate analysis for *PNPLA3* (6e-19 and 3e-9 for TAG 56:6;0 and cluster 1, respectively). TAG 56:6;0 is not contained in any multivariate cluster, which explains the higher *P*-value in the multivariate analysis.

New lipid-associated loci

Altogether, the univariate and multivariate GWAS identified 56 lipid-associated loci including 9 novel lipid loci (Table 2) in or near the following genes: *DTL*, *STK39*, *CDS1*, *AGPAT2*, *SGPL1*, *YPEL2*, *KCNJ12*, *SPHK2*, and *AGPAT3*. All these loci were identified by the multivariate GWAS but only two were also identified in the univariate GWAS (*AGPAT2* and *SGPL1*). The novel lead variants included missense variants for genes *AGPAT3* and *SPHK2*. *AGPAT2* and *AGPAT3* encode enzymes in the 1-acylglycerol-3-phosphate O-acyltransferase family, whose another member *AGPAT1* is known to be associated with standard lipids [17]–[20] and lipid species (PC, TAG) [26], [35]. AGPAT1/2/3 catalyze the conversion of lysophosphatidic acid to phosphatidic acid in the phospholipid and triacylglycerol synthesis. In our data, these regions were associated with PC and TAG species as well as cluster 3. MetaPhat analysis identified PC species to be driving the associations between cluster 3 and *AGPAT1/2/3* regions. *SPHK2*, associated with SM species and cluster 8, encodes a sphingosine kinase isozyme involved in sphingolipid metabolism. Three of the novel lead variants (*CDS1*, *SPHK2* and *STK39*) were > 2-fold enriched in Finland. The highest enrichment was 69-fold at the *CDS1* locus associated with cluster 2, with PI species as drivers of the association. Of note, *CDS1* encodes an enzyme that regulates the synthesis of PI.
We also report novel associations with lipid species for 11 loci which were previously identified in GWAS of standard lipids. These loci include *AC022784.1*, *ANKRD17*, *CPT1A*, *ELOVL6*, *ERMP1*, *GPAM*, *HNF4A*, *LCAT*, *MARC1*, *NPC1L1* and *PEX6* (Table 1, Supplementary Data 2).

Table 2. Novel loci reaching the Bonferroni corrected significance level for univariate or multivariate GWAS. Driver traits from metaPhat analysis are listed in parentheses after the cluster number for multivariate associations. Finnish enrichment is calculated as the ratio of minor allele frequencies (MAF) between our Finnish data and non-Finnish-non-Swedish-non-Estonian European samples in gnomAD v2.1 Enrichment values > 2 are bolded. The variant function is annotated by Variant Effect Predictor (VEP) and the gene naming the locus is identical to the gene annotated by VEP.

Locus	Trait	Lead variant	*P*-value	Function	MAF	Finnish enrichm.
DTL	cluster 5 (PC 18:0;0_18:2;0)	1:212081294:G:A	4e-11	intron	0.04	0.87
STK39	cluster 8 (SM 34:2;2)	2:168292271:A:G	3e-10	intergenic	0.03	**2.94**
CDS1	cluster 2 (PI 16:0;0_18:2;0)	4:84647685:C:T	2e-9	intron	0.07	**68.87**
AGPAT2	PC16:0;0_22:5;0 cluster 3 (PC 16:0;0_22:5;0, PC 16:0;0_22:4;0, PC 18:0;0_22:5;0)	9:136677616:C:G	4e-12 2e-12	intron	0.35	1.00
SGPL1	Cer42:2;2 cluster 8 (Cer 42:2;2,SM 34:2;2)	10:70843134:T:C	2e-10 6e-17	intron	0.21	0.86
KCNJ12	cluster 7 (PC 16:1;0_20:4;0)	17:21386711:C:T	2e-9	intron	0.38	0.99
YPEL2	cluster 11 (PC O-16:0;0/20:4;0)	17:59341010:C:T	2e-9	intron	0.06	1.42
SPHK2	cluster 8 (SM 34:2;2,SM 38:2;2)	19:48629610:G:C	6e-16	missense	0.03	**2.45**
AGPAT3	cluster 3 (PC 16:0;0_22:5;0, PC 18:0;0_22:5;0)	21:43971391:C:T	7e-17	missense	0.04	1.44

Fine-mapping of loci

To identify the most probable causal variants in the associated loci, we performed fine-mapping for both univariate and multivariate GWAS regions. Of the 56 loci, 26 loci had at least one variant with a high (> 0.9) posterior inclusion probability (PIP) in an informative 95% credible set either in univariate or multivariate GWAS (Supplementary Table 7). Altogether, 50 high PIP variants were identified. Variants with a high PIP were found from 13 loci in both univariate and multivariate analyses (*ABHD3*, *AGPAT2*, *APOE*, *CERS4*, *GCKR*, *GLTPD2*, *HNF4A*, *LINC01722*, *LIPC*, *PCSK9*, *PKD2L1*, *SMUG1P1*, and *ZPR1*), from 1 locus (*LPL*) only in univariate analysis and from 12 loci only in multivariate analysis (*AGPAT3*, *CPT1A*, *DOCK7*, *ELOVL6*, *LCAT*, *MYRF*, *NPC1L1*, *SGPL1*, *SMIM13*, *SPHK2*, *STK39*, *TM6SF2*). Of the 50 variants, 18 variants that reached a PIP > 0.9 in the multivariate analysis had a low PIP (< 0.1) in univariate analyses. In Supplementary Data 3 the full FINEMAP results are listed and the results for novel loci are summarized in Supplementary Table 6. Representative variants of informative credible sets of BFS univariate and multivariate GWAS regions are plotted by credible set size against top posterior inclusion probability (PIP) in Supplementary Figure 6.

In total, we found 53 variants that affect the molecular function of a protein among the representative variants of credible sets or in high LD (r^2 > 0.95) with them (Supplementary Data 3). These 53 functional variants were distributed among 32 of our 56 loci. For univariate analyses, 34 missense variants and 2 splice region variants were found across 24 loci. For multivariate analyses, 1 splice acceptor variant (*PNLIPRP2*), 2 splice donor variants (*LILRB3*, *ABHD12*), 1 frameshift variant (*ABHD12*), 1 inframe deletion variant (*NINL*), 38 missense variants and 2 splice region variants were found, distributed across 27 loci. Among the 18 functional variants with PIP > 0.5 in univariate or multivariate fine-mapping, 9 variants were predicted to be among the top 1% of most deleterious substitutions (CADD score > 20 [38]) and are reaching the GWS threshold in at least one GWAS. These are missense variants for genes *ABHD3*, *APOE*, *APOB*, *G6PC1*, *HNF4A*, *LCAT*, *LIPC*, *LPL*, and *SPHK2*.

Fine-mapping revealed multiple independent signals (Supplementary Data 2) at 24 of the 56 loci including novel signals for well-known lipid genes. For example, for *LIPC*, we found 5 signals represented by variants: rs6493996, rs2043085, rs1077834, rs113298164, and rs201563586, of which the last one is a highly Finnish-enriched missense variant, not in LD with any of the previously reported signals [20], [35], [39], [40].

A comparison to fine-mapping results of standard lipids in UKB was performed for 45 of our high PIP variants that were directly or through an LD-neighbor ($r^2 > 0.1$) contained in the UKB fine-mapping results (Supplementary Table 8). Of the 45 variants, 15 variants have a CADD score > 10, indicating that the variant is predicted to be among the 10 % of the most deleterious substitutions [38] (Table 3). Of these 15 variants, 8 variants reach a PIP > 0.1 in UKB for at least one standard lipid. The other 7 variants had a PIP < 0.001 for all standard lipids and were not contained in any 95 % credible set in UKB. Of the 7 variants, 3 variants were rare and only reached a PIP > 0.9 in our multivariate analysis. Detailed quality control assessment of these variants is in Supplementary Note. Of the 30 variants with a CADD score ≤ 10, 17 variants (or their LD neighbors) reach a PIP > 0.1 in UKB for at least one standard lipid. We observed a lower Pearson correlation between the standard lipids and lipid species or LCP-phenotypes associated with the variants that had only low PIP in UKB (Supplementary Table 8, Supplementary Note).

Table 3. Fine-mapping results. The table includes variants with a CADD score > 10 and a high PIP (> 0.9) in GeneRISK. Variants reaching only low PIP (< 0.1) in UKBB are marked with an asterisk. Function and gene are from VEP. Finnish enrichment is calculated as the ratio of minor allele frequencies (MAF) between Finnish samples and non-Finnish-non-Swedish-non-Estonian European samples in gnomAD v2.1 and listed if the variant is contained in gnomAD. Enrichment values > 2 are bolded. Traits for which the variant reaches a high PIP are listed and, in the case of multiple species of a lipid class, the number of species and the species for which the variant reaches the lowest *P*-value are given.

Locus	Variant	Function	Gene	CADD	Finnish enrichm.	MAF	Traits (*P*-value)
PCSK9	1:55039974:G:T	missense	*PCSK9*	10.4	**3.10**	0.033	CE 18:2;0 (2e-14), 3 SMs: SM 34:1;2 (2e-19), c3 (1e-14), c8 (2e-13)
GCKR	2:27508073:T:C	missense	*GCKR*	13.2	1.07	0.349	2 DAGs: DAG 18:1;0_18:2;0 (1e-12), 16 TAGs: TAG 50:4;0 (4e-22), c2 (4e-13), c3 (1e-17)
SMIM13	6:10995002:C:T*	splice_region	*ELOVL2*	22.9	**Inf**	0.004	c6 (2e-7)
LPL	8:19956018:A:G	missense	*LPL*	21.3	1.01	0.023	3 TAGs: TAG 54:4;0 (1e-9)
LIPC	15:58541944:G:A*	missense	*LIPC*	24.9	**Inf**	0.002	c2 (8e-8)
	15:58563549:C:T	missense	*LIPC*	24.1	**4.41**	0.017	5 PCs: PC 18:0;0_18:2;0 (1e-12), PC O-16:2;0/18:0;0 (3e-12), 5 PEs: PE 16:0;0_20:4;0 (4e-47), c2 (3e-62), c3 (2e-7), c4 (2e-10), c5 (4e-15), c7 (2e-7), c10 (1e-10)
LCAT	16:67942417:A:T	missense	*LCAT*	23.2	0.83	0.028	c7 (1e-16)
ABHD3	18:15528072:A:G*	intergenic		11.8		0.017	c2 (3e-6)
	18:21657147:C:T*	missense	*ABHD3*	23.7	**29.64**	0.004	c2 (4e-19), c3 (5e-12)
APOE	19:44908822:C:T	missense	*APOE*	26.0	0.56	0.053	5 CEs: CE 18:2;0 (2e-14), c3 (3e-23), c6 (2e-18), c7 (2e-53), c8 (2e-65), c11 (4e-14)
	19:44908684:T:C	missense	*APOE*	16.7	1.29	0.189	2 CEs: CE 20:2;0 (9e-12), c7 (1e-29), c9 (8e-12)
SPHK2	19:48629610:G:C*	missense	*SPHK2*	22.1	**2.45**	0.031	c8 (6e-16)
LINC01722	20:13160073:G:A*	missense	*SPTLC3*	18.0	**2.24**	0.086	3 Cers: Cer 42:2;2 (3e-17), c8 (7e-19)
HNF4A	20:44413724:C:T	missense	*HNF4A*	21.4	1.41	0.052	2 CEs: CE 18:3;0 (1e-10), c2 (1e-10)
AGPAT3	21:43971391:C:T*	missense	*AGPAT3*	16.3	1.44	0.039	c3 (7e-17)

Gene prioritization

Next, we prioritized genes in the 98 identified GWS loci first by using functional variants and second by using FOCUS [41], which together prioritized 49 genes (Supplementary Tables 9-11). First, we prioritized genes based on functional variants that had PIP > 0.5 or that were in high LD (r^2>0.95) with such variants. Of the 20 prioritized genes, 11 were found both in univariate and multivariate analysis (*AGPAT3*, *APOE*, *CERS4*, *CPT1A*, *GCKR*, *HNF4A*, *LIPC*, *PCSK9*, *SOAT2*, *SPTLC3*, *TM6SF2*), 3 only in univariate analysis (*G6PC1*, *LPL*, *TMC4*) and 6 only in multivariate analysis (*ABHD3*, *APOB*, *ELOVL2*, *ERMP1*, *LCAT*, *SPHK2*). FOCUS prioritized, at PIP > 0.5, 32 genes of which 17 were found both in univariate and multivariate analysis (*APOA5*, *AQP9*, *BFAR*, *CETP*, *CNOT3*, *DHX33*, *DOCK7*, *FNDC4*, *GRAMD4*, *LIPG*, *MIB1*, *NOMO1*, *PLEKHH1*, *PNPLA3*, *PPP6R1*, *SCGB2A2*, *SYNE2*), 4 only in univariate analysis (*APOB*, *APOA1*, *NLRP1*, *SCARB1*) and 11 only in multivariate analysis (*CCDC86*, *CERS4*, *CNN3*, *DDX49*, *ERMP1*, *GPAM*, *HNRNPM*, *MLEC*, *PRPF19*, *PYGB*, *ZNF506*).
We further assessed gene expression of the prioritized genes in 54 tissues using FUMA [42]. We observed high expression levels in liver for a majority of prioritized genes for both prioritization methods (Supplementary Figure 7). To assess tissue specificity of prioritized genes FUMA identifies differentially expressed genes (DEG) sets, defined as gene sets that are more (or less) expressed in a specific tissue compared to all other tissues. Up-regulated DEG sets were significantly enriched ($P \leq 0.05$ corrected for multiple testing) for liver tissue for both gene prioritization methods (Supplementary Figure 8). The top two enriched gene sets from Gene Ontology biological processes are 'lipid metabolic process' (adjusted P=3e-17) and 'cellular lipid metabolic process' (adjusted P=1e-15) for the functional variant approach, and 'protein containing complex remodeling' (adjusted P=1e-9) and 'lipid homeostasis' (adjusted P=2e-9) for FOCUS. The gene set enrichment results for the prioritized genes are in Supplementary Tables 12 and 13.

We assessed whether the prioritized genes were included in any gene set from FUMA with the name containing the term "lipid". For the functional approach, 3 out of 20 genes were not among the lipid gene sets (*ERMP1*, *G6PC1*, *TMC4*), and for FOCUS, the numbers were 20 out of 32 (e.g. *ZNF506*, *CNOT3*, *GRAMD4*). In total, of the 49 genes, 22 genes were not among FUMA's lipid gene sets.

Phenome-wide association study (PheWAS)

To explore the disease relevance of the identified lipid-associated loci, we used data for 953 disease endpoints from 377,277 participants from the FinnGen study. We performed PheWAS for the 264 GWAS lead variants and 287 representative variants of credible sets which were not among the lead variants. We identified 2,937 variant-disease associations for variants in 46 GWS loci reaching the *P*-value threshold of P < 5.25e-5 (corresponding to 0.05 corrected for the number of endpoints (953) included in the PheWAS; Supplementary Data 4). Amongst the 9 novel lipid-associated loci, PheWAS revealed an association at the locus *YPEL2* of the cluster 11 associated intronic variant 17:59341010:C:T with hypertension endpoints (minimum P=2e-7).

Figure 4 shows the connection between 9 selected PheWAS endpoints (representing cardiovascular disease, hyperlipidemia, diabetes, metabolic disorders, and neurological disease) and lipid species and multivariate clusters through common associated variants. Only the associations reaching the GWS threshold corrected for the number of endpoints (P < 5.25e-11 = 5e-8/953) are illustrated. Of the 179 lipid species, 137 species are included in Figure 4. We have listed the PheWAS associations of all endpoints and a list of endpoints included in each disease group in Supplementary Data 4. Supplementary Figure 9 shows the connection of all 45 BFS endpoints, ordered by disease groups, connected with > 3 lipid species.

Coronary artery disease loci associations

Of the 236 conditionally independent coronary artery disease **(**CAD) GWS variants at 196 loci [43], 11 reach the BFS threshold P < 7.35e-10 for univariate analysis and 1 additional variant reaches the BFS threshold P < 4.55e-9 for multivariate analyses (Figure 5). The most widely-associated variant is near *ZNF259*, located in the *BUD13-ZNF259-APOA5-APOA1-SIK3* gene-cluster, which increases the levels of DAGs, PCs, PE 18:0;0_18:2;0, PIs, and TAGs and decreases the level of PC O-16:1;0/18:1;0. This variant is also associated with statin medication (P=8e-130, Beta=+0.19) and disorders of lipoprotein metabolism and other lipidemias (P=9e-58, Beta=+0.18) in FinnGen R9. We summarized these associations and the 71 associations reaching the significance threshold corrected for multiple testing (P < 7.35e-4 for univariate and P < 4.55e-3 for multivariate analyses) in Supplementary Table 14. Of the 15 CAD variants that were GWS associated with a lipid species or clusters of lipid species, 4 variants with the nearest genes *NAT2* (rs4646249), *LPL* (rs268, rs894211), and *MYH11* (rs12691049) were not located within +/- 1.5 Mb of lipid variants reported by Cadby et al. [35] to be nominally associated with coronary atherosclerosis.

Discussion

We present a genetic study of plasma lipidome with 7,174 participants and 179 lipid species followed by a large-scale PheWAS analysis to reveal new lipid-associated variants and the relationship between lipid species and cardiometabolic disorders. Our study provided several advantages in gaining new information on the genetics of lipid metabolism due to (1) the large sample size of 7,174 individuals, (2) the unique genetic background of the Finnish population, (3) high resolution lipidomic measurements, and (4) the multivariate approach. We demonstrate a considerable gain of power from multivariate analysis of correlated lipid species compared to commonly used univariate analysis, and expand current knowledge in the field through the analysis of lipidome compared to the standard lipids. We identified variants that were highly associated with both lipid species and disease endpoints, including cardiovascular disease, liver disease, cholelithiasis, diabetes, and lipid disorders.

Our sample size is over 3-fold compared to the most recent GWAS on the same lipidome measures (2,181 individuals) (Tabassum et al. 2019 [21]). This increase is reflected in the number of univariate GWS findings (68 in our study vs. 35 in Tabassum et al.). Two other recent lipidome studies have been carried out with 5,662 Pakistani individuals plus 13,814 British individuals (Harshfield et al. 2021 [34]), and 4,492 Australian individuals predominantly of European ancestry (Cadby et al. 2022 [35]). Even though, the sample sizes in lipidome studies are still small compared to the existing GWAS on standard lipids ([19], [20]), high-dimensional lipidome phenotypes complement the standard lipid analyses by identifying new lipid-associated loci, providing a refined picture of the genetic associations and allowing multivariate analyses. Here, we have identified 15 lipid-associated loci that were not captured even by the largest GWAS of standard lipids with >1.65 million participants. The lipid species associated with these loci are less correlated with standard lipids than the remaining of the lipid species, reiterating that standard lipids do not completely capture the complex lipid metabolism.

The Finnish genetic background of our study population provides a unique opportunity to discover variants that are enriched in the Finnish population but extremely rare outside of Finland, and to identify new independent signals in known lipid loci. We identified three new lipid-associated loci that are enriched in the Finnish population, including a missense variant in *SPHK2* associated with SMs. *SPHK2* encodes sphingosine kinase 2 which plays an important role in sphingolipid metabolism. The unique LD pattern of the Finnish population also facilitated identification of additional independent variants associated with lipids in the known lipid loci through fine-mapping. For example, the *LIPC* region has been reported to contain three independent signals for standard lipids [20] and in addition to these a fourth independent signal has been reported to be associated with lipid species [35]. In addition to these four signals, our study identifies a new independent signal at a missense variant (rs201563586), not in LD with any of the previously reported signals. This variant

has a high PIP (> 0.90) in our fine-mapping analysis and is highly enriched in the Finnish population, indicating the benefits of studying population-isolates in genetic studies.

Another advantage of our study is the multivariate approach that showed a considerable gain in power in the discovery of new loci over the standard univariate GWAS. The multivariate GWAS identified 36% more BFS loci compared to univariate GWAS (55 vs. 35; Figure 3) and discovered 7 new loci (*DTL*, *STK39*, *CDS1*, *YPEL2*, *KCNJ12*, *SPHK2*, and *AGPAT3*), not detected by the univariate GWAS. The interpretation of a multivariate association is often not straightforward in terms of the original traits. Here we applied a recent statistical method [37] that decomposes the multivariate association into a smaller set of driver traits. Informative decompositions with only 2 or 3 driver traits were observed for 32% of the BFS GWAS regions (47 of 138) found by multivariate analysis. These regions represented eight such loci that did not reach BFS in any univariate analysis. Two examples are the association between cluster 7 and *APOB* locus driven by CE 16:0;0 and CE 20:4;0, and the novel association between cluster 8 and a missense variant in *SPHK2* driven by SM 34:2;2 and SM 38:2;2.

Individual lipid species have been shown to predict cardiovascular disease risk more accurately than standard lipids [21]. We observed disease associations with lipidome-associated variants for various disease groups (Figure 4). For statin medication, we observed a shared genetic association with 58% of the lipid species and all multivariate clusters. Another widely lipidome-associated endpoint was cholelithiasis (47%). The multivariate clusters and CE species are sharing genetic associations with all disease groups. Species of the classes SM, TAG, DAG, and a few species of the classes PC, PCO, PE, and PI show similar patterns for most disease groups, except for Alzheimer's disease, vascular dementia, or diabetic retinopathy. While these shared associations could point to interesting connections between lipid levels and diseases, there are two limitations with such observations. First, a shared association does not automatically mean that the potential causal variant in the region is the same for the lipid trait and the disease. Second, different disease endpoints in Figure 4 have varying effective sample sizes and therefore some differences between the observed associations across the diseases could simply reflect the differences in statistical power.

We also examined the lipidomic profiles of 11 known CAD variants (Figure 5). The CAD locus *ZNF259* showed the widest set of associations with 46 lipid species and 9 clusters of lipid species. The effect of the *ZNF259* polymorphism was previously only reported for standard lipids, with the first report [44] stating that individuals carrying the risk-increasing G allele showed increased TG levels and decreased LDL-C levels. We analyzed the effect of the polymorphism on lipid species: individuals with the G allele showed increased levels of DAGs, PCs, PE 18:0;0_18:2;0, PIs, and TAGs and decreased levels of PC O-16:1;0/18:1;0. Our list of marginal lipid associations of CAD-associated variants contained three such CAD loci associated at GWS with lipid species or clusters that were not included in the previous report [35] of CAD-associated lipid variants.

To summarize, our study identifies novel genetic loci with a role in lipid metabolism, points towards functional effects on detailed circulating lipid measures, and shows connections to cardio-metabolic and related diseases. We also highlight the benefits of utilizing multivariate methods for association testing in multiple correlated phenotypes. Our comprehensive catalog of detailed lipid associations provides new opportunities for studying the role of lipids in disease-associated loci.

Methods

Study participants

We use data from the prospective GeneRISK cohort whose principal aim is to assess the impact of communication of genetic risk information of CVD to study participants. The cohort includes 7,292 participants (4,642 women, 2,624 men), who were recruited from Southern Finland during 2015-2017 at age of 45-66 years. The sample collection and recruitment process are described in [45]. The basic study characteristics are presented in Supplementary Table 16. Participants were instructed to fast overnight for 10 hours before the collection of blood samples for plasma, serum, and DNA extraction. The biological samples (DNA, blood, serum, plasma) and the participants' demographic information and health data are stored in the THL Biobank (https://www.thl.fi/en/web/thlfien/topics/information-packages/thl-biobank). The GeneRISK study was carried out according to the principles of the Helsinki declaration and the Council of Europe's (COE) Convention of Human Rights and Biomedicine. All study participants gave their informed consent to participate in the study. The study protocols were approved by The Hospital District of Helsinki and Uusimaa Coordinating Ethics committees (approval No. 281/13/03/00/14 (GeneRISK)).

Ethics statement for FinnGen is listed in Supplementary Note.

Lipidomics

Mass spectrometry-based lipid analysis was performed for 7,302 individuals from the GeneRISK cohort by shotgun lipidomic analysis at Lipotype GmbH (Dresden, Germany). The analysis was performed by direct infusion in a QExactive mass spectrometer from Thermo Scientific with a TriVersa NanoMate ion source from Advion Biosciences [46]. The lipidomics data were analyzed using lipid identification software and a data management system developed in-house by Lipotype GmbH [47], [48]. Lipids with a high signal-to-noise ratio (> 5) and amounts at least 5-fold higher than corresponding blank samples were included. By including 8 reference samples per 96-well plate batch, reproducibility was assessed and lipid amounts were corrected for batch variations and analytical drift if the *P*-value of the slope was < 0.05 with an R^2 > 0.75 and the relative drift > 5%. Lipid species detected in more than 70% of the samples were included (179 lipid species from 13 lipid classes). After excluding samples with very low total lipid content and with > 30% of 179 lipids missing were excluded (N=26), data from 7,276 individuals remained.
Lipid molecules were identified at the species or subspecies level. Lipid species are named in the following notation: class name <sum of carbon atoms>:<sum of double bonds>;<sum of hydroxyl groups>. The annotation of lipid subspecies includes information on their acyl moieties and, if available, on their *sn*-position. The acyl chains are separated either by "_" if the *sn*-position on the glycerol cannot be resolved or else by "/". Further explanation is given by Gerl et al. [49]. Lipid identifiers of the SwissLipids database [50] (http://www.swisslipids.org) and the shorthand notation [51] are provided in Supplementary Table 2.

Genotyping and imputation

Genotyping was performed using the HumanCoreExome BeadChip from Illumina Inc. (San Diego, CA, USA) and genotype calling was done with GenomeStudio and zCall at the Institute for Molecular Medicine Finland (FIMM). Genotype data was lifted over to human genome build version 38 (GRCh38/hg38) according to the protocol described in dx.doi.org/10.17504/protocols.io.nqtddwn. In pre-imputation quality control (QC), potential outliers based on genetic ancestry were removed. We performed a principal component analysis (PCA) using 61,106 good quality (minor allele frequency (MAF) ≥ 0.05, Hardy-Weinberg equilibrium *P*-value (HWE) > 1e-6 and missingness < 10%) and approximately independent (LD pruning with PLINK v1.9: r^2 threshold of 0.2, window size 50 kb, step size 5) genetic variants. Based on the PCA and place of birth information from the questionnaire, individuals with non-Finnish ancestry or birthplace were removed. However, samples born in Estonia,

Russia, and Sweden, but clustered with the samples of Finnish ancestry in PCA, were retained in the analysis. Samples (N=30) with extreme heterozygosity (beyond ± 4 s.d) were excluded. After quality control filtering, 7,174 samples, consisting of 4,579 females and 2,595 males, with both genotype and lipidome data were considered for subsequent analyses.

Genotype data pre-phasing was done with Eagle 2.3.5 [52] with the number of conditioning haplotypes set to 20,000. Genotypes were imputed with Beagle 4.1 [53] (procedure described in https://doi.org/10.17504/protocols.io.nmndc5e) using population-specific Sequencing Initiative Suomi (SISu) v3 reference panel based on high-coverage (25–30x) whole-genome sequences for 3,775 Finnish individuals. In post-imputation QC, variants with imputation INFO score < 0.70 and MAF < 0.01 were excluded and 12,776,997 variants remained. The measured levels of the lipid species were adjusted for age, sex, collection site (clinic), lipid medication, first 10 genetic PCs, and ancestry (separate indicator variables for individuals born in Russia, Estonia, and Sweden) using linear regression. After adjusting for the above-mentioned covariates, the residuals were inverse-normal transformed and were used as outcome variables in the association analyses.

Hierarchical clustering of lipid species

Hierarchical clustering was performed using absolute pairwise Pearson correlations of plasma levels of lipids to identify clusters of correlated lipids for multivariate GWAS. The clustering analysis was performed separately for glycerolipids (44 lipid species from TAGs and DAGs) and the remaining lipid species (135 species belonging to glycerophospholipid, sphingolipid, and sterol). As highly correlated traits cause instability in multivariate association analyses, we iteratively excluded one member from each pair of lipid species with a correlation > 0.8 until no pair with a correlation > 0.8 remained. The hierarchical clustering was performed on the remaining lipid species using an average Euclidean distance metric on the pairwise correlations and clusters were identified by visually inspecting the dendrogram.
We then calculated Variance inflation factors (VIF) within each cluster for each cluster member using the R package ‘car’. (Technically, to apply the ‘car’ package, the cluster members were considered independent variables in a regression model where the outcome variable was a randomly generated variable whose exact value made no difference to the calculation of VIFs.) Through this approach, we identified cluster members that were highly correlated with some linear combination of the other members of the cluster. We iteratively removed the cluster member with the largest VIF until the maximum VIF within the cluster was below 5.

Hierarchical clustering of absolute pairwise Pearson correlations led to 11 clusters of correlated lipid species (Supplementary Figures 10 and 11). Based on the VIFs, one trait was removed from clusters 1 and 5 and two traits were removed from clusters 3 and 8. A heatmap of correlations for species included in the clusters is shown in Figure 1. A list of lipid species included in each cluster before and after removing traits is given in Supplementary Tables 17 and 1. Separate heatmaps of the correlations within each cluster are included in Supplementary Figure 12 and correlation values between lipid species are listed in Supplementary Table 18.

We computed pairwise Pearson correlations between the 179 lipid species and the standard lipids (HDL-C, LDL-C, TC, and TG). The correlation values are shown in Supplementary Figures 13 and 14 for lipid species and lipid classes, respectively. The correlation values between lipid species and standard lipids are listed in Supplementary Table 18. We obtained the maximum absolute correlation *maxCor* with any standard lipid for each lipid species and then calculated the *mean(maxCor)* for lipid species associated with loci not previously reported by standard lipids and for other species, to assess if lipid species associated with loci not reported by standard lipids are less correlated with standard lipids compared to other lipid species.

SNP-based heritability estimates

SNP-based heritability estimates for each lipid species were calculated using biMM [54]. The genetic relationship matrix (GRM) used for heritability estimates was calculated using 849,501 LD-pruned autosomal SNPs with imputation INFO score > 0.95, MAF > 0.01, and missingness < 3%. LD-pruning was done with PLINK v1.9 using a window size of 1000 kb, step size of 1, and pairwise r^2 threshold of 0.7. Additionally, high LD regions were excluded [55].

Univariate GWAS for 179 lipid species

Inverse-normal transformed residuals adjusted for the covariates mentioned above were used in the association analysis performed with the linear mixed model software MMM [56]. The number of samples per GWAS ranged between 5,287 and 7,174 because samples with missing values for a specific lipid species were excluded in the GWAS for that lipid species. After excluding very rare variants (MAF < 0.002) and variants with low imputation quality (INFO < 0.8), 11,318,730 variants were included in the GWAS. To account for multiple tests, the Bonferroni-corrected significance (BFS) threshold was set as *P*-value < 7.35e-10 (5e-8/68) as 68 principal components of the mean imputed lipidome data were required to explain > 90% of the phenotypic variance. All *P*-values reported in this study are two-sided.

Multivariate GWAS for 11 lipid clusters

Multivariate analysis of the 11 clusters identified through hierarchical clustering was performed with metaCCA [57].
Phenotypic correlations needed for the analysis were estimated from the GWAS summary statistics using metaCCA. Beta coefficients of the univariate GWAS were standardized using the formula suggested by metaCCA: $\beta_{stand} = \frac{\beta}{\sqrt{N}\, se}$, where *N* denotes the sample size of the respective univariate GWAS and *se* denotes the standard error. MetaCCA *P*-values were calculated from chi-square distribution using the mean GWAS sample size as parameter *N*. The BFS threshold for the multivariate GWAS was set to *P*-value < 4.55e-9 (5e-8/11) as the multivariate analysis was performed for 11 clusters.

The dataset used by Cichonska et al. [57] to test metaCCA consisted of variants with INFO > 0.8 and MAF > 0.05. To assess the robustness of the multivariate analysis for rare (MAF < 0.01) and low-frequency variants ($0.01 \leq$ MAF < 0.05) we simulated data under the null hypothesis for four SNPs with different MAFs covering range (0.005 – 0.042). In the simulation, the genotypes were permuted 100,000 times and then univariate GWAS were done with MMM and multivariate GWAS with metaCCA. The results of the simulation are described in detail in the Supplementary Note. We observed slightly inflated multivariate *P*-values for rare and low-frequency variants and are therefore correcting multivariate *P*-values of such variants using the genomic inflation factor (λ) [58] determined through this simulation approach. The simulation was performed for each rare or low-frequency variant that reached genome-wide significance level (P < 5e-8) in the multivariate analysis but not in any of the univariate analyses.

Further, for each of the clusters, MetaPhat [37] was applied to identify the traits driving the multivariate association at each lead variant of the multivariate analysis. The software determines sets of central traits for multivariate associations using Bayesian Information Criterion and *P*-value statistics. For each multivariate association, we report the driver traits and the optimal set of traits as defined by MetaPhat.

Defining lead variants in the GWAS regions

For both the univariate and multivariate GWAS, a lead variant in a GWAS was defined iteratively as the variant with the lowest *P*-value. After each new lead variant was identified, a 1.5 Mb region on each side of the variant defined the GWAS region of the lead variant, and other variants in that region were excluded from the further search for lead variants. For each GWAS, overlapping GWAS regions were combined into a single combined GWAS region, for which the lead variant is defined as the variant with the lowest *P*-value among the lead variants of the overlapping regions, and the other lead variants are listed as secondary lead variants. The maximum region width was set to 6 Mb, and for overlapping regions exceeding this threshold the original window size of +/- 1.5 Mb was iteratively shrunk by 10% until the width of the combined GWAS region was below 6 Mb (or the shrunk regions did not overlap anymore). The process was stopped after no variant outside the GWAS regions had reached genome-wide significance (GWS) of *P*-value < 5e-8. Similarly, we also defined the lead variants that reached Bonferroni-corrected significance (BFS).
To determine which of the lead variants from the multivariate analysis were also identified by the univariate analyses, we checked whether there were such variants that reached BFS or GWS in the univariate GWAS of any trait included in the multivariate analysis and had $r^2 > 0.1$ with the lead variant of the multivariate analysis.

A lead variant was considered "novel" if the lead variant was not in LD ($r^2 <0.1$) with any of the known variants identified in previous GWAS that included standard lipids or lipid species (listed in Supplementary Data 5). LD-proxies for previously reported variants that were not included in our GWAS were obtained using LDproxy from LDlink release 5.3.3 [59]. In LDproxy, we used the data on the 1000 Genomes project's Finnish population for LD calculation. For variants that were mono-allelic in the Finnish reference panel, we did the LD calculation with the combined European population. From the SNPs with $r^2 > 0.8$ and within 500 kb of the target variant, the one with the highest r^2 was chosen as LD-proxy. For 448 variants no proxy was found, of which 151 variants were monoallelic in both the Finnish and the European populations or were not biallelic variants, or were not contained in dbSNP build 155. For the remaining 297 of the 448 variants, none of the proxies were contained in our GWAS, no proxies with $r^2 > 0.8$ were found or the variants were not included in the 1000G reference panel. For these 297 previously reported variants, we additionally checked if any of our lead variants were located within +/- 1.5 Mb.

We report the closest gene for all lead variants using SNP-nexus [60]–[64] (overlapped gene if available or nearest upstream or downstream gene). The variant's function was predicted with Variant Effect Predictor (McLaren et al., 2016) and the most severe function was annotated to the variant. Possible functions were ordered by severity according to the ranking from Ensembl (https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html). We defined functional variants as having at least one of the following functions (ordered by severity from more severe to less severe): transcript_ablation, splice_acceptor_variant, splice_donor_variant, stop_gained, frameshift_variant, stop_lost, start_lost, transcript_amplification, inframe_insertion, inframe_deletion missense_variant, protein_altering_variant, splice_region_variant.

GWAS of linear combination phenotypes (LCP-GWAS)

To enable fine-mapping of multivariate associations, linear combination phenotypes (LCP) were constructed as a weighted sum of the traits where the weights corresponded to the optimal combination phenotype reported by metaCCA for the lead variant [65]. GWAS region-specific LCP-GWAS were performed with fastGWA-mlm [66], with the same covariates as were used in the univariate GWAS.

We calculated pairwise Pearson correlations between the LCP phenotypes and the standard lipids.

(Supplementary Data 2).

Fine-mapping

Fine-mapping was performed with FINEMAP [67] for each GWAS region. For each fine-mapped region, the in-sample linkage disequilibrium (LD) matrix was computed using LDstore2 [68] from genotype dosages. The maximum number of causal variants in a locus was set to 10. The number of independent association signals for each fine-mapped GWAS region was determined by the number of informative credible sets (CS) among those CS for which FINEMAP gave the highest posterior probability. CS was considered informative if the minimum r^2 among its variants was ≥ 0.1. We chose the top variant from each CS to represent the association signal except if the CS contained functional variants in high LD ($r^2 > 0.95$) with the top variant, in which case the functional variant having the largest r^2 with the top variant was chosen as the representative variant [65]. The GWAS lead variant was chosen as the representative variant if no informative CS was obtained. The MHC region (chr 6: 25 Mb - 34 Mb) was excluded from the fine-mapping and there the GWAS lead variant was defined as the representative variant.

Defining independent signals and physical loci across all traits

Earlier we defined GWAS regions separately in each univariate or multivariate GWAS and these regions were used in fine-mapping. Next, we used the representative variants from the fine-mapping results to determine the set of independent signals across all traits. We merged the signals across the traits if their representative variants were in LD ($r^2 \geq 0.1$). For each signal, we took the union of the corresponding GWAS regions to define physical boundaries for the signal and finally we combined the overlapping signal regions to form a single set of physical loci across all traits. The locus definition process is summarized in a flow chart and visualized for the locus LPL in Supplementary Figures 3 and 4, respectively. Locus names were defined by the closest gene to the top variant with the lowest *P*-value across the associated traits except if there was a missense variant among the top variants, in which case the locus was named by the gene corresponding to the missense variant. Novel loci are defined as loci containing only GWAS regions whose lead variants were all novel.

Comparison of fine-mapping results to fine-mapping results of standard lipids

We checked how the variants that got a high posterior inclusion probability (PIP) > 0.9 in the GeneRISK data, or other variants in the same locus in LD with them ($r^2 > 0.1$ in GeneRISK), were fine-mapped across the standard lipids (HDL-C, LDL-C, TG, TC) in the UK Biobank (UKB) data by Finucane lab (https://www.finucanelab.org/data). For this, the chromosomal positions of the UKB data were lifted over to human genome build version 38 (GRCh38/hg38) with liftOver [69]. We considered only the variants included in both data sets. We acknowledge that the UKB variants that were not present in GeneRISK, but that were in LD with a GeneRISK variant with a high PIP, could explain why some high PIP variants in GeneRISK may have lower PIP in UKB.

Gene prioritization and pathway enrichment analysis

We prioritized genes for which we found functional variants with PIP > 0.5 in fine-mapping of the univariate or multivariate GWAS. For the functional variants, we obtained functional variant scores from Variant Annotation Integrator [70] and CADD scores from CADD v1.6 [38].

Additionally, we performed a gene prioritization analysis using FOCUS [41], which computes credible sets of genes based on a posterior inclusion probability (PIP). We performed Transcriptome-wide association studies (TWAS) and tissue-agnostic fine-mapping with FOCUS using GTEx v8 eQTL reference panel weight database and in-sample LD. We used MASHR-based GTEx v8 eQTL databases from PrediXcan [71]–[73] to create the weight database. As input, we used univariate GWAS and multivariate LCP-GWAS filtered for INFO > 0.8 and MAF > 0.002 and cleaned data with

the munge command from FOCUS. We classified the GTEx tissues into two categories, category 1 containing subcutaneous adipose tissue, visceral adipose tissue, liver, and whole blood, which were deemed most relevant for lipid-related phenotypes in a previous study [34], and category 2 containing the remaining tissues.

We utilized FUMA software's GENE2FUNC tool [42] to obtain information on the expression of the prioritized genes and identify pathways enriched for the prioritized genes.

Phenome-wide association analyses

Phenome-wide association analyses (PheWAS) were performed for the GWAS lead variants and the representative variants of credible sets in 377,277 participants from the FinnGen biobank (FinnGen release 9) [74]. From FinnGen, all 953 endpoints of the following categories (ICD-10 Chapter listed in parentheses if available) were included: 'cardiometabolic endpoints', 'diabetes endpoints', 'diseases marked as autoimmune origin', 'drug purchase endpoints', 'gastrointestinal endpoints', 'neoplasms from hospital discharge' (II), 'neoplasms from cancer register' (II), 'diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism' (III), 'endocrine nutritional and metabolic diseases' (IV), 'diseases of the nervous system' (VI), 'diseases of the circulatory system' (IX), 'neurological endpoints', 'diseases of the digestive system' (XI). These ICD-10 chapters were chosen because diseases within these chapters have been previously reported to be associated with changes in lipid metabolism, such as II: breast cancer [75], III: Systemic Lupus Erythematosus [76], IV: lipid metabolism disorders and diabetes mellitus [77], VI: Alzheimer's disease [78], IX: Coronary artery disease [35], XI: Nonalcoholic Fatty Liver Disease [79]. For all endpoints at least 50 cases exist. The included endpoints for each data source are listed in Supplementary Data 4. We report associated endpoints reaching the threshold $P < 0.05$ corrected for the number of included endpoints ($P < 0.05/953 = 5.25e\text{-}5$) for each lead variant and representative variants of credible sets. Additionally, we identified endpoints reaching the GWS threshold corrected for the number of included endpoints ($P < 5e\text{-}8/953 = 5.25e\text{-}11$). Due to the high correlation between many endpoints these thresholds might be too stringent.
We focused on PheWAS endpoints connected with > 3 lipid species or multivariate clusters and then assigned disease groups to endpoints. We selected 11 endpoints of 5 disease groups to be included in a heatmap. The selected endpoints were chosen by selecting the endpoint with the largest effective sample size N_{eff} among endpoints of the same disease and by selecting endpoints with the most specific diagnoses based on expert medical knowledge. N_{eff} was defined as $N\,\theta(1-\theta)$, with θ being the proportion of cases. We provide a list of the endpoints and their disease groups and effective sample size in Supplementary Data 4, where the selected endpoints are highlighted.

Association of coronary artery disease loci with lipidome

We assessed associations of the coronary artery disease (CAD) variants identified by a recent study [43] with lipid species and clusters of lipid species in our study. Of the 241 conditionally independent GWS associations with CAD at 198 loci, 236 variants at 196 loci were either included in our GWAS, or their LD-proxies were found in our GWAS (LD proxies were defined using the same approach as with the lead variants). We summarized the associations at three levels of significance: (1) $P < 0.05$ corrected for multiple testing by the number of PCs explaining 90% of the variance (univariate analyses) or the number of clusters (multivariate analyses) ($P < 0.05/68 = 7.35e\text{-}4$ for univariate and $P < 0.05/11 = 4.55e\text{-}3$ for multivariate analyses), (2) the GWS threshold $P < 5e\text{-}8$ and (3) the GWS threshold corrected for multiple testing ($P < 5e\text{-}8/68 = 7.35e\text{-}10$ for univariate and $P < 5e\text{-}8/11 = 4.55e\text{-}9$ for multivariate analyses).

Data availability

Univariate GWAS summary statistics will be available on GWAS catalog (https://www.ebi.ac.uk/gwas/) upon publication.
DNA, blood, serum, and plasma samples of GeneRISK study participants, in addition to their demographic information and health data, are stored in the THL Biobank (https://thl.fi/en/web/thl-biobank/).

References

[1] T. Vos *et al.*, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, Oct. 2020, doi: 10.1016/S0140-6736(20)30925-9.

[2] A. Wienke, A. M. Herskind, K. Christensen, A. Skytthe, and A. I. Yashin, "The heritability of CHD mortality in danish twins after controlling for smoking and BMI," *Twin Res Hum Genet*, vol. 8, no. 1, pp. 53–59, Feb. 2005, doi: 10.1375/1832427053435328.

[3] J. Borén *et al.*, "Low-density lipoproteins cause atherosclerotic cardiovascular disease: pathophysiological, genetic, and therapeutic insights: a consensus statement from the European Atherosclerosis Society Consensus Panel," *European Heart Journal*, vol. 41, no. 24, pp. 2313–2330, Jun. 2020, doi: 10.1093/eurheartj/ehz962.

[4] R. Tabassum and S. Ripatti, "Integrating lipidomics and genomics: emerging tools to understand cardiovascular diseases," *Cell Mol Life Sci*, vol. 78, no. 6, pp. 2565–2584, Mar. 2021, doi: 10.1007/s00018-020-03715-4.

[5] C. Stegemann *et al.*, "Lipidomics profiling and risk of cardiovascular disease in the prospective population-based Bruneck study," *Circulation*, vol. 129, no. 18, pp. 1821–1831, May 2014, doi: 10.1161/CIRCULATIONAHA.113.002500.

[6] Z. H. Alshehry *et al.*, "Plasma Lipidomic Profiles Improve on Traditional Risk Factors for the Prediction of Cardiovascular Events in Type 2 Diabetes Mellitus," *Circulation*, vol. 134, no. 21, pp. 1637–1650, Nov. 2016, doi: 10.1161/CIRCULATIONAHA.116.023233.

[7] R. Laaksonen *et al.*, "Plasma ceramides predict cardiovascular death in patients with stable coronary artery disease and acute coronary syndromes beyond LDL-cholesterol," *Eur Heart J*, vol. 37, no. 25, pp. 1967–1976, Jul. 2016, doi: 10.1093/eurheartj/ehw148.

[8] A. S. Havulinna *et al.*, "Circulating Ceramides Predict Cardiovascular Outcomes in the Population-Based FINRISK 2002 Cohort," *Arterioscler Thromb Vasc Biol*, vol. 36, no. 12, pp. 2424–2430, Dec. 2016, doi: 10.1161/ATVBAHA.116.307497.

[9] L. R. Peterson *et al.*, "Ceramide Remodeling and Risk of Cardiovascular Events and Mortality," *J Am Heart Assoc*, vol. 7, no. 10, p. e007931, May 2018, doi: 10.1161/JAHA.117.007931.

[10] J. W. Meeusen, L. J. Donato, S. C. Bryant, L. M. Baudhuin, P. B. Berger, and A. S. Jaffe, "Plasma Ceramides," *Arterioscler Thromb Vasc Biol*, vol. 38, no. 8, pp. 1933–1939, Aug. 2018, doi: 10.1161/ATVBAHA.118.311199.

[11] P. A. Mundra *et al.*, "Large-scale plasma lipidomic profiling identifies lipids that predict cardiovascular events in secondary prevention," *JCI Insight*, vol. 3, no. 17, pp. e121326, 121326, Sep. 2018, doi: 10.1172/jci.insight.121326.

[12] C. Razquin *et al.*, "Plasma lipidome patterns associated with cardiovascular risk in the PREDIMED trial: A case-cohort study," *Int J Cardiol*, vol. 253, pp. 126–132, Feb. 2018, doi: 10.1016/j.ijcard.2017.10.026.

[13] M. Hilvo *et al.*, "Development and validation of a ceramide- and phospholipid-based cardiovascular risk estimation score for coronary artery disease patients," *Eur Heart J*, vol. 41, no. 3, pp. 371–380, Jan. 2020, doi: 10.1093/eurheartj/ehz387.

[14] A. M. Poss *et al.*, "Machine learning reveals serum sphingolipids as cholesterol-independent biomarkers of coronary artery disease," *J Clin Invest*, vol. 130, no. 3, pp. 1363–1376, Mar. 2020, doi: 10.1172/JCI131838.

[15] T. M. Teslovich *et al.*, "Biological, clinical and population relevance of 95 loci for blood lipids," *Nature*, vol. 466, no. 7307, pp. 707–713, Aug. 2010, doi: 10.1038/nature09270.

[16] C. J. Willer *et al.*, "Discovery and refinement of loci associated with lipid levels," *Nat Genet*, vol. 45, no. 11, pp. 1274–1283, Nov. 2013, doi: 10.1038/ng.2797.

[17] I. Surakka *et al.*, "The impact of low-frequency and rare variants on lipid levels," *Nat Genet*, vol. 47, no. 6, pp. 589–597, Jun. 2015, doi: 10.1038/ng.3300.

[18] D. J. Liu *et al.*, "Exome-wide association study of plasma lipids in >300,000 individuals," *Nat Genet*, vol. 49, no. 12, pp. 1758–1766, Dec. 2017, doi: 10.1038/ng.3977.

[19] D. Klarin *et al.*, "Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program," *Nat Genet*, vol. 50, no. 11, pp. 1514–1523, Nov. 2018, doi: 10.1038/s41588-018-0222-9.

[20] S. E. Graham *et al.*, "The power of genetic diversity in genome-wide association studies of lipids," *Nature*, vol. 600, no. 7890, pp. 675–679, Dec. 2021, doi: 10.1038/s41586-021-04064-3.
[21] R. Tabassum *et al.*, "Genetic architecture of human plasma lipidome and its link to cardiovascular disease," *Nat Commun*, vol. 10, no. 1, p. 4329, Sep. 2019, doi: 10.1038/s41467-019-11954-8.
[22] C. Gieger *et al.*, "Genetics Meets Metabolomics: A Genome-Wide Association Study of Metabolite Profiles in Human Serum," *PLOS Genetics*, vol. 4, no. 11, p. e1000282, Nov. 2008, doi: 10.1371/journal.pgen.1000282.
[23] A. A. Hicks *et al.*, "Genetic Determinants of Circulating Sphingolipid Concentrations in European Populations," *PLOS Genetics*, vol. 5, no. 10, p. e1000672, Oct. 2009, doi: 10.1371/journal.pgen.1000672.
[24] T. Illig *et al.*, "A genome-wide perspective of genetic variation in human metabolism," *Nat Genet*, vol. 42, no. 2, pp. 137–141, Feb. 2010, doi: 10.1038/ng.507.
[25] K. Suhre *et al.*, "Human metabolic individuality in biomedical and pharmaceutical research," *Nature*, vol. 477, no. 7362, Art. no. 7362, Sep. 2011, doi: 10.1038/nature10354.
[26] A. Demirkan *et al.*, "Genome-Wide Association Study Identifies Novel Loci Associated with Circulating Phospho- and Sphingolipid Concentrations," *PLOS Genetics*, vol. 8, no. 2, p. e1002490, Feb. 2012, doi: 10.1371/journal.pgen.1002490.
[27] E. P. Rhee *et al.*, "A genome-wide association study of the human metabolome in a community-based cohort," *Cell Metab*, vol. 18, no. 1, pp. 130–143, Jul. 2013, doi: 10.1016/j.cmet.2013.06.013.
[28] S.-Y. Shin *et al.*, "An atlas of genetic influences on human blood metabolites," *Nat Genet*, vol. 46, no. 6, pp. 543–550, Jun. 2014, doi: 10.1038/ng.2982.
[29] H. H. M. Draisma *et al.*, "Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels," *Nat Commun*, vol. 6, no. 1, Art. no. 1, Jun. 2015, doi: 10.1038/ncomms8208.
[30] T. Long *et al.*, "Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites," *Nat Genet*, vol. 49, no. 4, pp. 568–578, Apr. 2017, doi: 10.1038/ng.3809.
[31] A. Demirkan *et al.*, "Genome-wide association study of plasma triglycerides, phospholipids and relation to cardio-metabolic risk factors." bioRxiv, p. 621334, Jun. 22, 2019. doi: 10.1101/621334.
[32] L. A. Lotta *et al.*, "A cross-platform approach identifies genetic regulators of human metabolism and health," *Nat Genet*, vol. 53, no. 1, pp. 54–64, Jan. 2021, doi: 10.1038/s41588-020-00751-5.
[33] K. A. McGurk *et al.*, "Heritability and family-based GWAS analyses of the N-acyl ethanolamine and ceramide plasma lipidome," *Human Molecular Genetics*, vol. 30, no. 6, pp. 500–513, Mar. 2021, doi: 10.1093/hmg/ddab002.
[34] E. L. Harshfield *et al.*, "Genome-wide analysis of blood lipid metabolites in over 5000 South Asians reveals biological insights at cardiometabolic disease loci," *BMC Medicine*, vol. 19, no. 1, p. 232, Sep. 2021, doi: 10.1186/s12916-021-02087-1.
[35] G. Cadby *et al.*, "Comprehensive genetic analysis of the human lipidome identifies loci associated with lipid homeostasis with links to coronary artery disease," *Nat Commun*, vol. 13, no. 1, Art. no. 1, Jun. 2022, doi: 10.1038/s41467-022-30875-7.
[36] P. Helkkula *et al.*, "ANGPTL8 protein-truncating variant associated with lower serum triglycerides and risk of coronary disease," *PLoS Genet*, vol. 17, no. 4, p. e1009501, Apr. 2021, doi: 10.1371/journal.pgen.1009501.
[37] J. Lin, R. Tabassum, S. Ripatti, and M. Pirinen, "MetaPhat: Detecting and Decomposing Multivariate Associations From Univariate Genome-Wide Association Statistics," *Front Genet*, vol. 11, p. 431, 2020, doi: 10.3389/fgene.2020.00431.
[38] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Research*, vol. 47, no. D1, pp. D886–D894, Jan. 2019, doi: 10.1093/nar/gky1016.
[39] T. Tukiainen *et al.*, "Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci," *Human Molecular Genetics*, vol. 21, no. 6, pp. 1444–1455, Mar. 2012, doi: 10.1093/hmg/ddr581.

[40] A. Gallois *et al.*, "A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context," *Nat Commun*, vol. 10, no. 1, p. 4788, Oct. 2019, doi: 10.1038/s41467-019-12703-7.
[41] N. Mancuso *et al.*, "Probabilistic fine-mapping of transcriptome-wide association studies," *Nat Genet*, vol. 51, no. 4, pp. 675–682, Apr. 2019, doi: 10.1038/s41588-019-0367-1.
[42] K. Watanabe, E. Taskesen, A. van Bochoven, and D. Posthuma, "Functional mapping and annotation of genetic associations with FUMA," *Nat Commun*, vol. 8, no. 1, p. 1826, Nov. 2017, doi: 10.1038/s41467-017-01261-5.
[43] K. G. Aragam *et al.*, "Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants," *Nat Genet*, vol. 54, no. 12, Art. no. 12, Dec. 2022, doi: 10.1038/s41588-022-01233-6.
[44] S. Kathiresan *et al.*, "Common variants at 30 loci contribute to polygenic dyslipidemia," *Nat Genet*, vol. 41, no. 1, pp. 56–65, Jan. 2009, doi: 10.1038/ng.291.
[45] E. Widén *et al.*, "How Communicating Polygenic and Clinical Risk for Atherosclerotic Cardiovascular Disease Impacts Health Behavior: an Observational Follow-up Study," *Circulation: Genomic and Precision Medicine*, vol. 0, no. 0, p. CIRCGEN.121.003459, doi: 10.1161/CIRCGEN.121.003459.
[46] M. A. Surma *et al.*, "An automated shotgun lipidomics platform for high throughput, comprehensive, and quantitative analysis of blood plasma intact lipids," *Eur J Lipid Sci Technol*, vol. 117, no. 10, pp. 1540–1549, Oct. 2015, doi: 10.1002/ejlt.201500145.
[47] R. Herzog *et al.*, "A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language," *Genome Biology*, vol. 12, no. 1, p. R8, Jan. 2011, doi: 10.1186/gb-2011-12-1-r8.
[48] R. Herzog *et al.*, "LipidXplorer: a software for consensual cross-platform lipidomics," *PLoS One*, vol. 7, no. 1, p. e29851, 2012, doi: 10.1371/journal.pone.0029851.
[49] M. J. Gerl *et al.*, "Machine learning of human plasma lipidomes for obesity estimation in a large population cohort," *PLoS Biol*, vol. 17, no. 10, p. e3000443, Oct. 2019, doi: 10.1371/journal.pbio.3000443.
[50] L. Aimo *et al.*, "The SwissLipids knowledgebase for lipid biology," *Bioinformatics*, vol. 31, no. 17, pp. 2860–2866, Sep. 2015, doi: 10.1093/bioinformatics/btv285.
[51] G. Liebisch *et al.*, "Update on LIPID MAPS classification, nomenclature, and shorthand notation for MS-derived lipid structures," *J Lipid Res*, vol. 61, no. 12, pp. 1539–1555, Dec. 2020, doi: 10.1194/jlr.S120001025.
[52] P.-R. Loh *et al.*, "Reference-based phasing using the Haplotype Reference Consortium panel," *Nat Genet*, vol. 48, no. 11, Art. no. 11, Nov. 2016, doi: 10.1038/ng.3679.
[53] B. L. Browning and S. R. Browning, "Genotype Imputation with Millions of Reference Samples," *Am J Hum Genet*, vol. 98, no. 1, pp. 116–126, Jan. 2016, doi: 10.1016/j.ajhg.2015.11.020.
[54] M. Pirinen, C. Benner, P. Marttinen, M.-R. Järvelin, M. A. Rivas, and S. Ripatti, "biMM: efficient estimation of genetic variances and covariances for cohorts with high-dimensional phenotype measurements," *Bioinformatics*, vol. 33, no. 15, pp. 2405–2407, Aug. 2017, doi: 10.1093/bioinformatics/btx166.
[55] A. L. Price *et al.*, "Long-Range LD Can Confound Genome Scans in Admixed Populations," *Am J Hum Genet*, vol. 83, no. 1, pp. 132–135, Jul. 2008, doi: 10.1016/j.ajhg.2008.06.005.
[56] M. Pirinen, P. Donnelly, and C. C. A. Spencer, "Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies," *The Annals of Applied Statistics*, vol. 7, no. 1, pp. 369–390, Mar. 2013, doi: 10.1214/12-AOAS586.
[57] A. Cichonska *et al.*, "metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis," *Bioinformatics*, vol. 32, no. 13, pp. 1981–1989, Jul. 2016, doi: 10.1093/bioinformatics/btw052.
[58] B. Devlin and K. Roeder, "Genomic control for association studies," *Biometrics*, vol. 55, no. 4, pp. 997–1004, Dec. 1999, doi: 10.1111/j.0006-341x.1999.00997.x.
[59] M. J. Machiela and S. J. Chanock, "LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants," *Bioinformatics*, vol. 31, no. 21, pp. 3555–3557, Nov. 2015, doi: 10.1093/bioinformatics/btv402.

[60] C. Chelala, A. Khan, and N. R. Lemoine, "SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms," *Bioinformatics*, vol. 25, no. 5, pp. 655–661, Mar. 2009, doi: 10.1093/bioinformatics/btn653.

[61] A. Z. Dayem Ullah, N. R. Lemoine, and C. Chelala, "SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update)," *Nucleic Acids Research*, vol. 40, no. W1, pp. W65–W70, Jul. 2012, doi: 10.1093/nar/gks364.

[62] A. Z. Dayem Ullah, N. R. Lemoine, and C. Chelala, "A practical guide for the functional annotation of genetic variations using SNPnexus," *Briefings in Bioinformatics*, vol. 14, no. 4, pp. 437–447, Jul. 2013, doi: 10.1093/bib/bbt004.

[63] A. Z. Dayem Ullah, J. Oscanoa, J. Wang, A. Nagano, N. R. Lemoine, and C. Chelala, "SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine," *Nucleic Acids Research*, vol. 46, no. W1, pp. W109–W113, Jul. 2018, doi: 10.1093/nar/gky399.

[64] J. Oscanoa, L. Sivapalan, E. Gadaleta, A. Z. Dayem Ullah, N. R. Lemoine, and C. Chelala, "SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update)," *Nucleic Acids Research*, vol. 48, no. W1, pp. W185–W192, Jul. 2020, doi: 10.1093/nar/gkaa420.

[65] S. E. Ruotsalainen *et al.*, "An expanded analysis framework for multivariate GWAS connects inflammatory biomarkers to functional variants and disease," *Eur J Hum Genet*, vol. 29, no. 2, pp. 309–324, Feb. 2021, doi: 10.1038/s41431-020-00730-8.

[66] L. Jiang, Z. Zheng, H. Fang, and J. Yang, "A generalized linear mixed model association tool for biobank-scale data," *Nat Genet*, vol. 53, no. 11, pp. 1616–1621, Nov. 2021, doi: 10.1038/s41588-021-00954-4.

[67] C. Benner, C. C. A. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, and M. Pirinen, "FINEMAP: efficient variable selection using summary data from genome-wide association studies," *Bioinformatics*, vol. 32, no. 10, pp. 1493–1501, May 2016, doi: 10.1093/bioinformatics/btw018.

[68] C. Benner, A. S. Havulinna, M.-R. Järvelin, V. Salomaa, S. Ripatti, and M. Pirinen, "Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies," *The American Journal of Human Genetics*, vol. 101, no. 4, pp. 539–551, Oct. 2017, doi: 10.1016/j.ajhg.2017.08.012.

[69] R. M. Kuhn, D. Haussler, and W. J. Kent, "The UCSC genome browser and associated tools," *Brief Bioinform*, vol. 14, no. 2, pp. 144–161, Mar. 2013, doi: 10.1093/bib/bbs038.

[70] A. S. Hinrichs *et al.*, "UCSC Data Integrator and Variant Annotation Integrator," *Bioinformatics*, vol. 32, no. 9, pp. 1430–1432, May 2016, doi: 10.1093/bioinformatics/btv766.

[71] A. Barbeira *et al.*, "Fine-mapping and QTL tissue-sharing information improves the reliability of causal gene identification," *Genetic epidemiology*, vol. 44, Sep. 2020, doi: 10.1002/gepi.22346.

[72] A. N. Barbeira, M. Pividori, J. Zheng, H. E. Wheeler, D. L. Nicolae, and H. K. Im, "Integrating predicted transcriptome from multiple tissues improves association detection," *PLOS Genetics*, vol. 15, no. 1, p. e1007889, Jan. 2019, doi: 10.1371/journal.pgen.1007889.

[73] E. R. Gamazon *et al.*, "A gene-based association method for mapping traits using reference transcriptome data," *Nat Genet*, vol. 47, no. 9, pp. 1091–1098, Sep. 2015, doi: 10.1038/ng.3367.

[74] M. I. Kurki *et al.*, "FinnGen: Unique genetic insights from combining isolated population and national health register data." medRxiv, p. 2022.03.03.22271360, Mar. 06, 2022. doi: 10.1101/2022.03.03.22271360.

[75] N. Marino *et al.*, "Upregulation of lipid metabolism genes in the breast prior to cancer diagnosis," *npj Breast Cancer*, vol. 6, no. 1, Art. no. 1, Oct. 2020, doi: 10.1038/s41523-020-00191-8.

[76] W. Sun *et al.*, "Lipid Metabolism: Immune Regulation and Therapeutic Prospectives in Systemic Lupus Erythematosus," *Frontiers in Immunology*, vol. 13, 2022, Accessed: Aug. 30, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2022.860586

[77] V. Natesan and S.-J. Kim, "Lipid Metabolism, Disorders and Therapeutic Drugs – Review," *Biomol Ther (Seoul)*, vol. 29, no. 6, pp. 596–604, Nov. 2021, doi: 10.4062/biomolther.2021.122.

[78] H. Chew, V. A. Solomon, and A. N. Fonteh, "Involvement of Lipids in Alzheimer's Disease Pathology and Potential Therapies," *Frontiers in Physiology*, vol. 11, 2020, Accessed: Aug. 30, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphys.2020.00598

[79] K. Pei *et al.*, "An Overview of Lipid Metabolism and Nonalcoholic Fatty Liver Disease," *Biomed Res Int*, vol. 2020, p. 4020249, Jul. 2020, doi: 10.1155/2020/4020249.

Acknowledgements

We would like to thank Johanna Aro, Sari Kivikko, and Ulla Tuomainen for management assistance in the project. We thank all study participants of the GeneRISK study for their participation. We acknowledge the participants and investigators of the FinnGen study. Full FinnGen acknowledgments and FinnGen funders are provided in the supplemental acknowledgments in the Supplementary Note. The GeneRISK study was funded by Business Finland through the Personalized Diagnostics and Care program coordinated by SalWe Ltd (grant No. 3986/31/2013). Dr Ripatti was supported by the Academy of Finland Center of Excellence in Complex Disease Genetics (grant No. 312062), the Finnish Foundation for Cardiovascular Research, the Sigrid Juselius Foundation, and University of Helsinki HiLIFE Fellow and Grand Challenge grants. Dr Pirinen was supported by the Academy of Finland (grants 338507 and 336825) and Sigrid Juselius Foundation.

Ethics declarations

Competing interests

K.S. is CEO of Lipotype GmbH. K.S. and C.K. are shareholders of Lipotype GmbH. M.J.G. is employee of Lipotype GmbH.

Author information

Contributions

L.O., R.T., M.J.G., K.S., S.R. and M.P. conceived and designed the study; L.O. performed multivariate GWAS and all statistical analyses and reported the results; R.T. performed univariate GWAS; S.E.R. performed quality control of genotype data; L.O., R.T., M.J.G. E.W., K.S., S.R. and M.P. interpreted the results; M.J.G., C.K. and K.S. performed lipidomic profiling and processed the raw data; L.O. drafted the manuscript with help from R.T. and M.P; R.T., S.R. and M.P. supervised the study. All authors read, commented, and approved the manuscript.

[6] Full author list in Supplementary Table 19

Figures

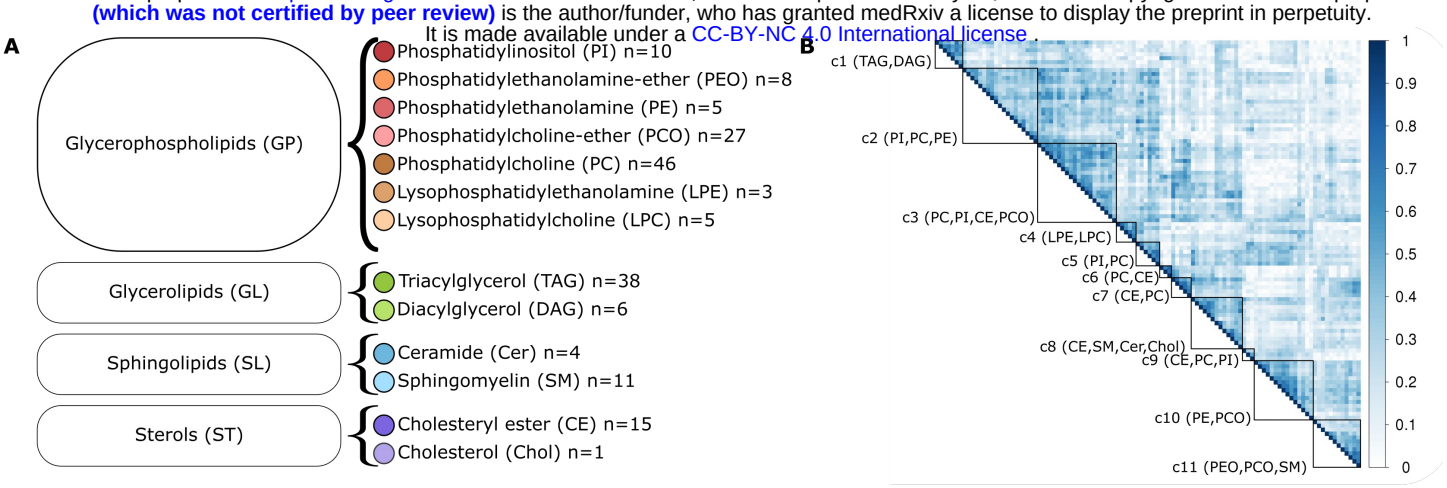

Figure 1. Details of lipid species measured in the GeneRISK cohort. (A) The 179 lipid species belong to 13 lipid classes and 4 categories. Lipid class colors are identical to those used in other figures. **(B) Heatmap of absolute pairwise Pearson correlations between lipid species included in the 11 clusters of the multivariate GWAS.** Clusters are marked by black squares and labeled by lipid classes. The members of each cluster are listed in Supplementary Table 1.

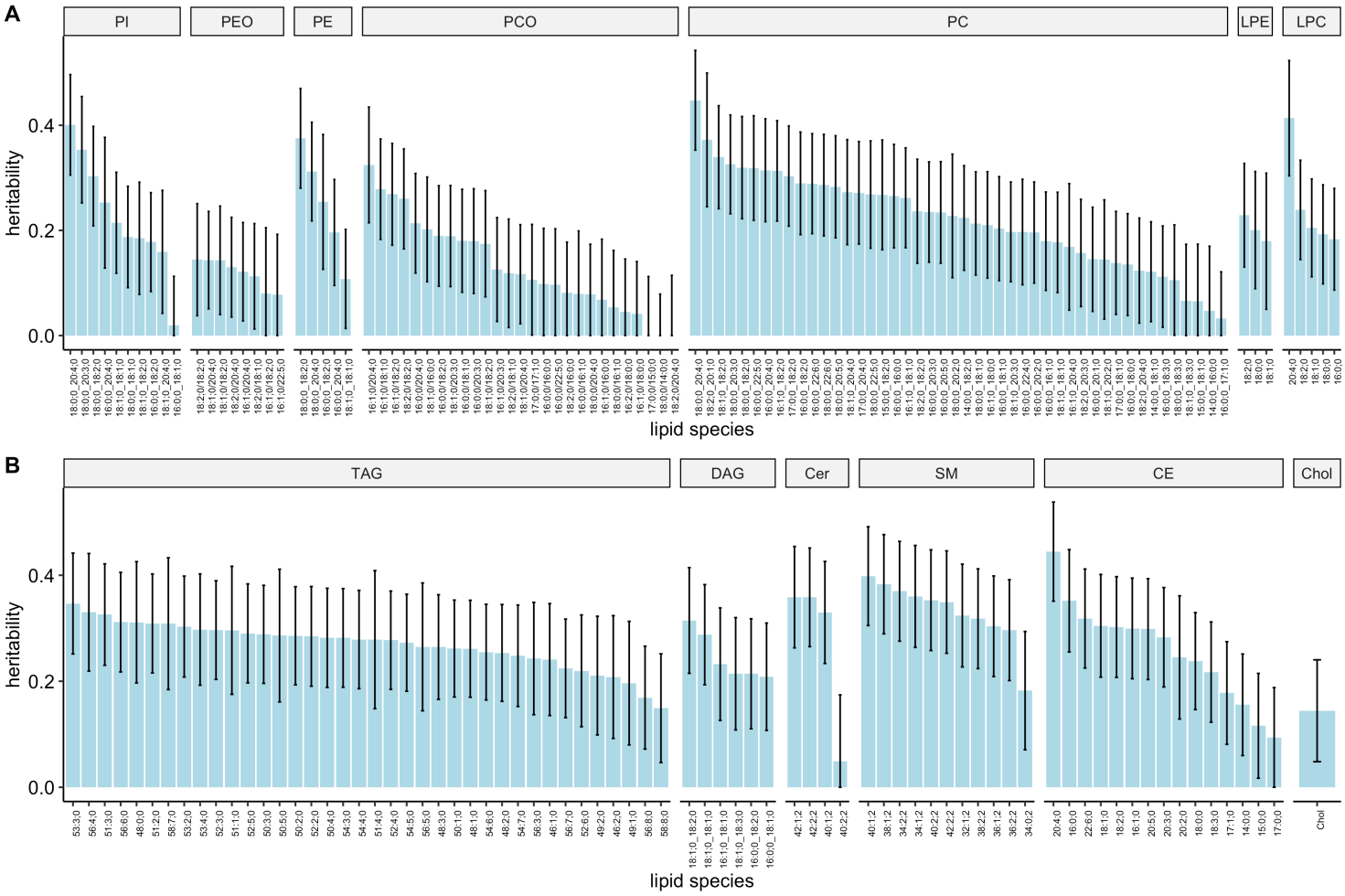

Figure 2. Heritability estimates of lipid species. Panel A shows Glycerophospholipids and panel B shows Glycerolipids, Sphingolipids, and Sterols. Error bars represent 95% confidence intervals. Lipid species are presented in the descending order of the heritability estimates.

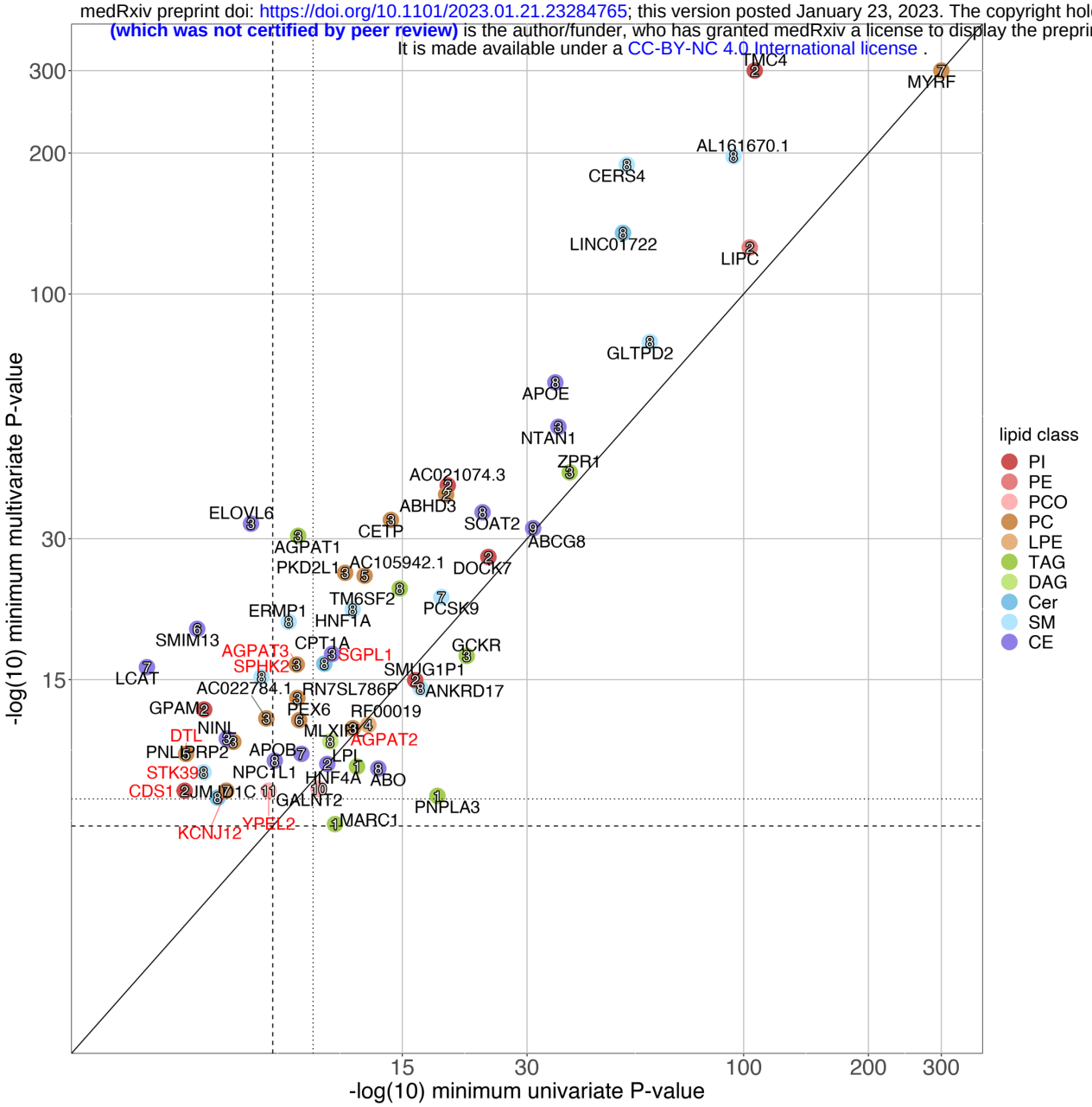

Figure 3. Comparison of the univariate and multivariate *P*-values for 56 lipid-associated loci. Loci are colored by lipid class in univariate analysis and labeled by cluster number from multivariate analysis. X-axis shows the *P*-values of the top associated univariate lead variant of the loci. Y-axis shows the *P*-values of the top associated multivariate lead variant of the loci. If no variant reached $P < 5e\text{-}8$ for the locus in univariate analysis, the minimum univariate *P*-value of the lead variant of the multivariate analysis is shown. Known loci and novel loci are annotated by locus name in black and red, respectively. Dashed lines represent the genome-wide significance level ($P < 5e\text{-}8$) and dotted lines represent the Bonferroni-corrected significance level (uv: 7.35e-10, mv: 4.55e-9). Lipid class names are listed in Figure 1. Axes are capped at 300.

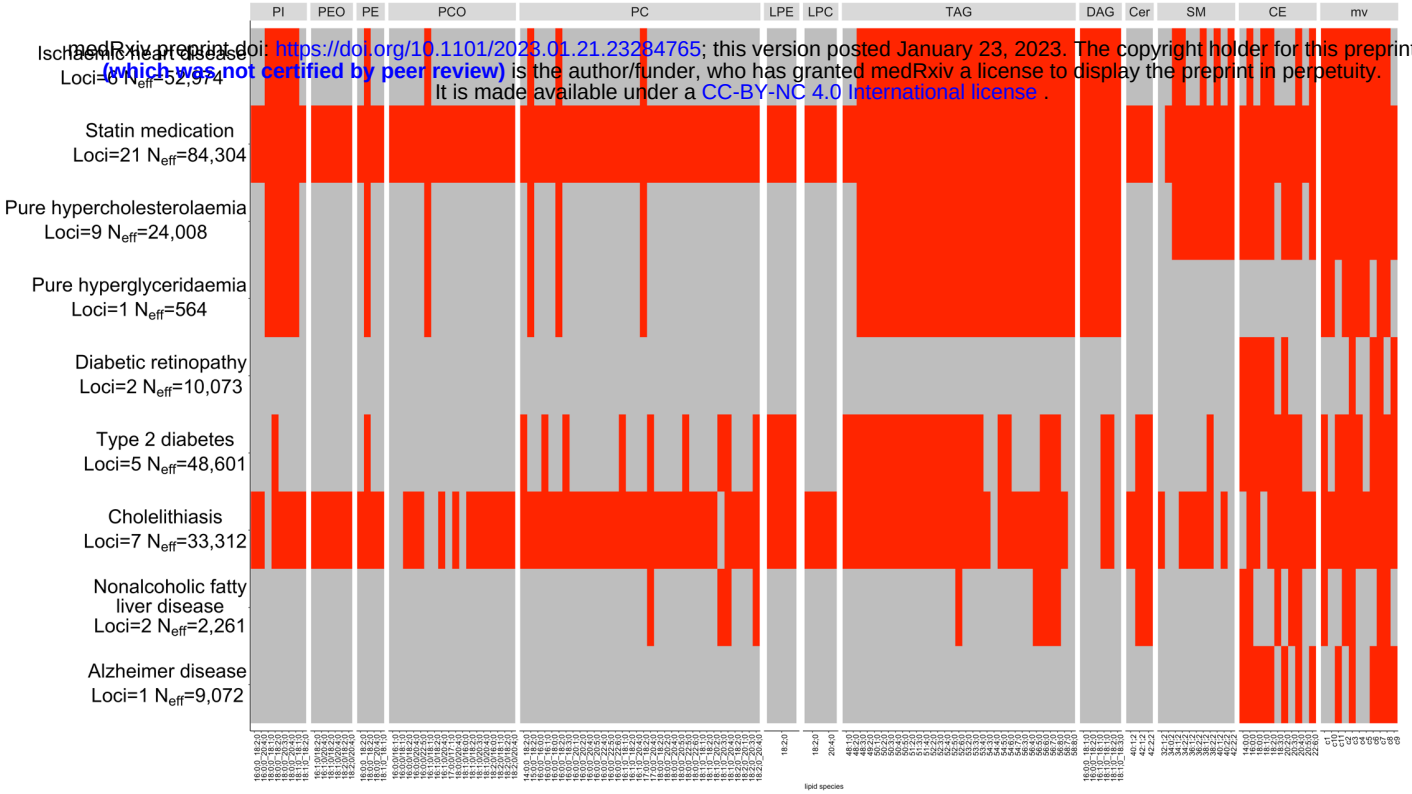

Figure 4. Heatmap of PheWAS associations for selected disease endpoints. Each entry in the heatmap represents a possible association between a disease group (row) and a lipidome trait (column). Red color indicates that at least one variant among the lead variants or representative variants of the lipidome trait is also associated with the disease at P < 5.25e-11. Gray denotes that no such association is observed. Columns are split by lipid classes. The effective sample size N_{eff} (see Methods) and the number of loci are given beneath each disease endpoint.

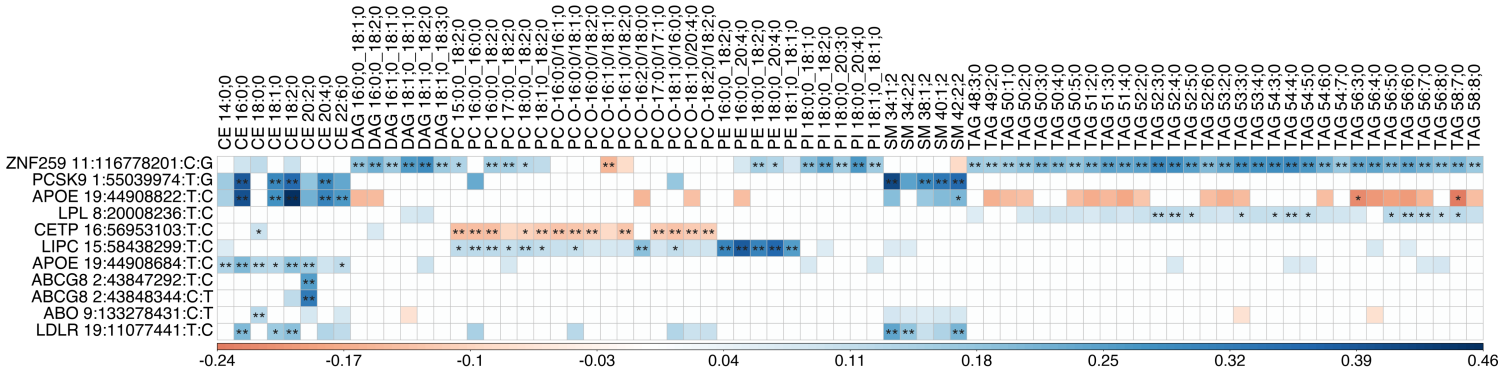

Figure 5. Effect estimates of 11 CAD risk-increasing alleles on lipid species. Variant ids are defined as Chromosome:Position:risk-decreasing allele:risk-increasing allele. Included are species that reach the BFS threshold of 7.35e-10 for at least one of the variants. Associations reaching GWS (P < 5e-8) or BFS are indicated by one or two asterisks, respectively. Colored effect estimates are shown for associations reaching nominal significance corrected for the number of PCs explaining 90% of the variance (P < 7.35e-4).